09/647875

PCT/CN 99/00046

CN99/46

证　明

5

本证明之附件是向本局提交的下列专利申请副本

申　　请　　日：98 04 06

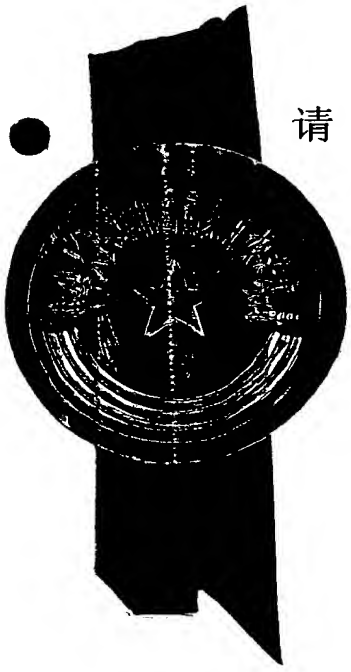
申　　请　　号：98 1 01156.X

申 请 类 别：发　明

发明创造名称：全息全选全程模板式人机对话
语言翻译方法

发明人或设计人：刘　莎

请　　人：刘　莎



中华人民共和国
国家知识产权局局长 姜颖

99 年 04 月 09 日

权　利　要　求　书

1.一种全息全选全程模板式人机对话语言翻译方法。其特征在于包括下述步骤:

a.建立一个以句子为对象的包括各种自然语言必要信息要素的自然语言翻译人机对话模板;

b. 由对话模板提供对不同自然语言进行统一通约受限后的所有备选信息项;

c.先由翻译系统对统一通约受限后的所有备选信息项进行自动优选，再由用户在全息对话模板上对优选结果进行人工调整和确认;

d.由翻译系统根据确定信息项进行译出目标语转换生成，将源语输入方的选项结果随译文提供给用户查询。

2.根据权利要求 1 所述的全息全选全程模板式人机对话语言翻译方法。其特征在于: 所述步骤 b 的不同自然语言统一通约受限方法是对基础概念进行强制性对齐，不能用基础概念进行统一的自然语言词汇或概念，在对话模板中提供空白信息项。

3.根据权利要求 1 所述的全息全选全程模板式人机对话语言翻译方法。其特征在于: 所述步骤 a 的所有必要信息要素包括由概念定义、时态信息及语态信息组成的普通信息项和句法信息项。

4. 根据权利要求 1 所述的全息全选全程模板式人机对话语言翻译方法。其特征在于: 所述步骤 c 的对自动优选结果进行人工调整和确认的方法是由用户在全息对话模板上对不确定信息进行人工选择。

5.根据权利要求 1 所述的全息全选全程模板式人机对话语言翻译方法。其特征在于: 所述步骤 a 的以句子为对象的人机对话模板是包括有三维空间定位句法的对话框架。

6.根据权利要求 1 所述的全息全选全程模板式人机对话语言翻译方法。其特征在于: 所述步骤 a 的以句子为对象的人机对话模板是虚拟的。

7.根据权利要求 1 所述的全息全选全程模板式人机对话语言翻译方法。其特征在于所述步骤 b 的不同自然语言统一通约受限包括: a. 统一合并功能同一、对象不同一的句法概念; b. 尽量删除可缺少的句法概念; c. 通过对主要语种词汇使用频率的统计分析和同义归并，建立多语通用基础概念; d.以各种自然语言的基础概念近义词作为近义附码，当不同自然语言出现近义词对应空缺时，由基础概念词进行近义替换; e.不能用基础概念进行统一表达的自然语言词汇或概念，由对话模板提供空白信息项; f.对话模板中提供用户选择的是经简化通约后的信息项。

8.根据权利要求 1 所述的全息全选全程模板式人机对话语言翻译方法。其特征在于：不同自然语言统一通约受限方法还包括有词汇概念通约，是 a .以内涵为中心的模糊通约和 b． 不考虑词性差异的概念统一通约。

9.根据权利要求 1 所述的全息全选全程模板式人机对话语言翻译方法。其特征在于：所述步骤 c 中，用户可单项或多项地在全息对话模板上对优选结果进行人式调整和确认选择。

说　明　书

全息全选全程模板式人机对话语言翻译方法

本发明涉及一种计算机翻译方法，更确切地说是涉及一种在计算机网络中适于各网络终端以不同自然语言进行信息传递交流的机器翻译方法。

计算机网络技术以其四通八达、无处不到的优势而迅速开创出一个全球化的网络信息时代。但由于不同自然语言之间语义信息的传递交流障碍，已明显制约了网络及网络信息的使用效率，如何通过机器翻译处理使各网络终端用户仅使用自己的自然语言在网络上进行语义信息传递，对于节省网络空间、提高网络信息的传递效率和实现网络信息资源的大众化国际共享，都无疑具有重要的现实意义和很高的商业价值。

目前在机器翻译领域，一方面由人工智能教科书上系统介绍的机器翻译方法在实际的产品开发中很少被使用，另一方面，在已开发出的机器中所应用的机器翻译方法又不能达到预期的目标，上述现象说明：基础理论研究严重滞后；所有的机器翻译技术方法都具有普遍共性的缺陷；预期目标本身不具有现实性。进入90年代以来，出现了大致两类新兴的机器翻译方法并逐渐成为自然语言信息处理的技术主流。一种是以对大规模真实文本的统计分析为基本手段建设语料库，另一种是人机对话及自然语言受限的机器翻译方法。

大规模真实文本的统计分析是通过对大规模真实文本进行符号、句型、词性、语义等多角度的信息取样分析，从而为任何一种自然语言中的符号串提供多种匹配模式，因而是一种基于经验的语言信息处理方法。从方法论上说用这种自然语言信息处理方法所获得的统计结果具有客观真实性及良好的可应用性，但从应用角度分析，这种语言信息处理方法仍然是一种提供匹配模式的方法，与传统的句型模式匹配方法无本质区别。理论上是可以将源语的多种匹配分析结果作叠加处理并通过与译出目标语的多种匹配分析结果建立匹配关系，而直接完成自然语言的自动翻译，但现实状况是，自然语言系统具有随机开放特性，任何统计方法都只能提供概率性知识，不可能对自然语言词汇及其概念定义进行准入限制，不可能确定各种省略表达部分的确切内容，也不可能解决生成目标语后的新增歧义。因此，大规模真实文本的统计分析虽然对于利用计算机进行各种自然语言信息处理来说确是有意义的基础工作，但对于机器翻译来说，这种技术手段还需要组合在一种全面有效的对象处理系统方法中才能充分实现其应用价值。

人机对话及自然语言受限的机器翻译方法也有传统及新型两种技术方案。传统方法包括由用户在输入端调整机器翻译词典和调整源语言表达方式，同时调整译文结果，该方法虽可获得较好的机器翻译质量，但要求用户熟练掌握机器翻译

的源语和目标语并需付出相当高的人机对话学习成本及操作成本，与人工翻译不差上下。新型方案的人机对话方案只要求用户熟练掌握母语和学会规范表达，适应机器翻译系统给出的源语表达规范，以满足机器翻译在源语分析方面的要求，但即使是规范的自然语言表达也仍然存在一词多义和句法歧义，其判别还需借助语境语义分析，因此仅依靠建立源语表达规范，是不能真正解决源语信息求解中的所有问题的。即使通过人机对话使受限的自然语言机器翻译系统完成源语信息求解的任务，但如果不能有效解决目标语生成后的新增歧义问题，是难以保证机器翻译系统的翻译质量的。

本发明的目的是设计一种全息全选全程模板式人机对话机器翻译方法，以全面解决计算机网络多语种信息传递交流障碍问题，试图取得机器翻译技术的实质性突破。这种突破必须满足以下条件:

1．对自然语言普通词汇及其概念定义进行有效的准入限制;

2．不依赖上下文语境进行语义分析;

3．通过直译手段实现语义信息准确传递;

4．找到生成目标语后的新增歧义解决办法;

5．用户只需熟练掌握母语;

6．利用大规模真实文本统计分析的手段与成果，充分实现人机优势互补;

7．满足向多种目标语言转换的需要。

本发明的目的是这样实现的，全息全选全程人机对话机器翻译方法，其特征在于包括下述步骤:

a .建立一个以句子为对象的包括各种自然语言必要信息要素的自然语言翻译人机对话模板;

b．由对话模板提供对不同自然语言进行统一通约受限后的所有备选信息项;

c .先由翻译系统对统一通约受限后的所有备选信息顶进行自动优选，再由用户在全息对话模板上对优选结果进行人工调整和确认;

d .由翻译系统根据确定信息项进行译出目标语转换生成，将源语输入方的选项结果随译文提供给用户查询。

所述步骤 b 的不同自然语言统一通约受限方法是对基础概念进行强制性对齐，不能用基础概念进行统一的自然语言词汇或概念，在对话模板中提供空白信息项。

所述步骤 a 的所有必要信息要素包括由概念定义、时态信息及语态信息组成的普通信息项和句法信息项。

所述步骤 c 的对自动优选结果进行人工调整和确认的方法是由用户在全息对话模板上对不确定信息进行人工选择。

所述步骤 a 的以句子为对象的人机对话模板是包括有三维空间定位句法的对

话框架。

所述步骤 a 的以句子为对象的人机对话模板是虚拟的。

所述步骤 b 的不同自然语言统一通约受限包括：1．统一合并功能同一、对象不同一的句法概念；2．尽量删除可缺少的句法概念；3．通过对主要语种词汇使用频率的统计分析和同义归并，建立多语通用基础概念；4.以各种自然语言的基础概念近义词作为近义附码，当不同自然语言出现近义词对应空缺时，由基础概念词进行近义替换；5.不能用基础概念进行统一表达的自然语言词汇或概念，由对话模板提供空白信息项；6.对话模板中提供用户选择的是经简化通约后的信息项。

不同自然语言统一通约受限方法还包括有词汇概念通约，是 1.以内涵为中心的模糊通约和 2．不考虑词性差异的概念统一通约。

所述步骤 c 中，用户可单项或多项地在全息对话模板上对优选结果进行人式调整和确认选择。

本发明全息全选全程模板式人机对话机器翻译方法的技术特点是：人机对话的基本点是由用户对模板信息直接进行选择，对用户而言只需掌握母语，基本无学习成本；本方法是在充分考虑计算机对信息处理的实际边界能力并以语义信息传递的准确性为中心任务及实际目标而作出的；本方法充分利用了人机优势互补，翻译内容不受语言环境和应用领域限制；本方法通过建立统一受限标准和全息全选全程的人机对话，提供了一揽子解决机器翻译基本技术障碍的系统方案，为根本改善机器翻译质量提供了全方位的技术保证；本方法可充分利用大规模语料库建设的成果，对自然语言简洁实用的处理方法，使其具有良好的可实施性；虽然在源语信息求解阶段，用户看不懂的语言不可能进行人机对话，但可在保证翻译质量的前提下实现一种语言输入得到多语种译出结果。

本发明的全息全选全程模板式人机对话机器翻译方法在网络信息交流领域具有普遍应用的价值，在打开网络在线机器翻译服务方面有广阔的国际市场。

图 1 是以句子为对象的自然语言全息模型结构示意图

图 2-1 、 2-2 、 2-3 、 2-4 是四种全息全选全程对话模板结构示意图

图 3 是句法信息的空间定位结构示意图

图 4 是普通概念统一受限编码框架结构示意图

图 5 是源语选项结果查询结构示意图

下面结合一句子的中英全息全选全程模板式人机对话语言翻译实施过程进一步说明本发明的技术。该句子为“我在银行附近看见一个带望远镜的男孩”，“I saw a boy with a telescope near the bank.”

首先建立一个以句子为对象的自然语言全息对话模型，在这个模型中包括各种自然语言文字符号系统所有必要的语言信息要素，为所要进行的人机对话作自然语言词汇及其概念定义的有效准入限制。该模型如图 1 所示，以句子为对象的

人机对话模板是一种包括三椎空间定位句法的对话框架。

图 1 中所有必要的语言信息要素包括由概念义项信息项、时态信息项和语态信息项构成的普通信息项和由句法成分项构成的句法信息项。从机译词典中调出示例句子中各符号串的相应信息项内容并填入模型中，如图 2-1 中所示。

要根本改善机器翻译质量、提高机器翻译系统的实用价值，必须对模板对话信息项进行通约受限。

为了准确传递语义信息，最好采用直译手段，这是因为机器翻译系统不可能随机调整目标语句子的词汇和句型。但要想保证直译的译文质量，必须保证词汇信息项和句法信息项能在源语与目标语间作等价交换。因此本发明对不同自然语言间的差异通过建立系统的通约受限原则进行统一整合处理。这种通约受限原则包括句法信息通约和普通信息通约。

本发明设计的句法信息通约原则包括：统一合并功能同一、对象不同一的句法信息；尽量删除在语义聚合关系分析中并非不可缺少的句法概念，如英语语法中的直接宾语与间接宾语。本发明在对话模板上只提供经简化通约后的句法信息概念，作为不同自然语言的标准句法信息项供用户选择。

本发明设计的普通信息通约如图 4 中所示，是通过对大语种词汇使用频率的统计分析和同义归并而确定一个基础概念集。但实际操作时，不是每一种自然语言的基础概念都是完整的，当出现空缺时，则要采用该语言对这一概念进行解释性描述，使基础概念强制性对齐。如英文词汇 orphan 的动词义项被定为基础概念，而中文中没有对应词，则用“使成为孤儿”进行解释性描述。此外，一种自然语言中某个词汇的全部近义概念也不可能在其它自然语言中全部找到对应概念，因此在当某种自然语言的近义概念出现对应空缺时则由基础概念词进行近义替换(人工翻译中近义替换也是不可避免的)。经过上述两项通约处理后仍不能处理的则作为冗余信息在全息模型中提供空白信息项。本发明在确定不同自然语言词汇的概念定义时，采用以内涵为中心的模糊通约(如中文的“学校”与英文的”school”);不考虑词性差异的概念统一通约(如不考虑英文词汇 become 的所有时态变形)和对多种语言中都使用的概念作优先考虑的概率通约处理，为了丰富语言的表达力，任何语言都需要有同一概念的近义词，因此以词汇的使用概率作为普通概念冗余标准，优先多种语言中都使用的概念，其次是在一种自然语言使用概率高的词汇。对于不满足上述两种情况的词汇则作为冗余概念处理。如汉语中“看”的近义词“睃”、“内顾”、“谛视”等都作为冗余概念。经过通约受限处理后的词汇信息才作为全息模板中的词汇备选项提供给不同自然语言用户进行选择，以保证不同自然语言普通概念信息间能够等价互换。

本发明的对多种自然语言概念系统进行强制性通约受限的方法，与传统的中间语言方法间有着本质区别：传统的中间语言技术面对的是完全不受限的自然语言系统，通过建立多种自然语言间的中间概念体系来实现多语互译，但各种自然

语言概念体系的开放性使中间语言体系不可能具有周延性；强制性的通约受限方法是通过人机对话方式对词汇及义项作必要的限制和通约，对各种自然语言概念体系之间的差异和开放性进行合理限制，以保证多种自然语言的词汇概念及句法概念能成功地进行等价互换。

在普通信息项的选择中要充分利用人机优势互补，计算机自动优选所遵循的基本原则是：通过大规模的对真实文本的统计分析，排列出多义词的词汇信息项使用频率顺序，以缩小用户选项的搜寻范围；通过大规模的对真实文本的统计分析，根据句法信息项与词汇信息项间的相关性特性来优选词汇信息项，以进一步缩小信息项选择范围，如凡可做主语的词汇都优选其名词义项，图 2 中我、望远镜、银行等；通过大规模的对真实文本的统计分析，获得词汇搭配的概率信息，进一步优选词汇信息项，如汉语“好漂亮的一朵花”，其中的“好”是多义词，而在形容词“漂亮”前的“好”字的最可能的义项解是程度副词“非常”；对于显性表达词性信息的文字符号，通过词性即可推导出所选词汇信息项来缩小信息项选择范围，如英语中“ spring ”的词根虽然是多义的，但其动词的过去式“ sprang ”则已明确限制了义项选择范围。

通过以上技术手段的自动选项处理，已能够将用户实际所需的大多数词汇信息项排在首位，由于表达语义所需要的词汇信息项是在用户心中的，因此对用户而言，大多数的普通信息项选择只是一个对模型中各首选信息项的确认过程。

各种自然语言中，无论是隐性表达还是显性表达的句法信息，大体上包括词性信息、句法成分信息和上位语义信息，其中句法成分信息是唯一具有完整组织能力的，并具有普遍共性的句法组织系统，因此，只要确定句法成分信息项，实际上已经确定了一个自然语言符号串的语义聚合关系。在句法信息项的选择中也要充分利用人机优势互补，其所遵循的基本原则是：通过大规模的对真实文本的统计分析获得词序、词性、上位语义信息与句法信息之间的匹配关系，以自动优选句法信息项。如一个词汇的词序为 1，词性为名词，上位语义为行为主体，则可判定为主语；用户通过选项操作最终确定句法成分信息项。

通过全选式人机对话过程最后确定词汇信息项和句法信息项，求解自然语言的信息。由用户直接在全息对话模型上选择各自然语言符号串实际携带的词汇信息项和句法信息项，是最简单的人机对话方式，其具体方法可以是对所确定的项进行黑体标注处理，如图 2 中所示。

通过在全息模型中对句子中词汇信息项和句法信息项的人机互补选择、确认，已能够完成自然语言的信息求解任务，因此不再需要依赖上下文语境对句子进行语义分析，

对于用户来说，分析和确定抽象的句法关系远比判断多义词信息项困难，因此，为了降低句法成分信息项的选择难度，实际操作时可象图 3 中所示的那样将呈线性排列的句法成分信息项转换成空间定位表达方式，协助进行句法成分信息

项人机对话的选择。以句法信息的修饰区、核心区及补充区为横座标，以句法信息的主语区、谓语区及宾语区为纵座标，作出句法信息对话框架，由用户在框架中对“ with a telescope”的修饰对象进行选择。

在实际的人机对话过程中也可以采用模板部分显示方法和模板虚拟方法，如图 2-2 所示的句法信息全显(图中？号表示由用户再选择)，图 2-3 所示的词汇信息(带)的单个全选和图 2-4 所示的“ I see a boy with a telescope near the bank ”的虚拟对话模板后的对话显示方法

本发明的方法通过对语法概念和普通概念的系统通约受限，以及在受限信息项范围内进行人机互补信息全选，已经具有了向多种自然语言表达形式作自动转换的必要信息，但总有被用户省略的句法成分，从逻辑上说只要确定了已有文字符号的所有信息项，大多数省略部分可由用户在阅读信息时根据上下文语境自动添加(如主词、宾词省略)，但为了准确传递语义，对不可省略的句子成分还要通过全息对话模型进行添加，以保证机器翻译质量(如在一个句子的备选信息项中已经选了主词和宾词，则不可省略相关动词)。

参见图 5 ，图中示出在找到生成目标语后发现新增歧义的解决办法。将经过全息对话的中间翻译结果随译文提供给目标语用户作直接查询，可实现目标语新增歧义的全面消解。被查询信息的显示模板也可以采用图 2-2 、 2-3 、 2-4 所示的形式。如果用户有意保留语言表达的模糊性或双关性，则可在选择信息项时作多项同时选择。

语义信息传递质量是全球化网络信息时代机器翻译技术赢得巨大国际市场的根本障碍，要想取得实质性突破，人机对话是不可避免的，本发明人机对话优势互补的翻译方案可切实提高翻译质量，具有实用价值。由于本方法具有语言信息传递准确、不受语言环境限制、用户操作使用方便、可同步转换生成多种目标语、对话方案多语通用及技术手段简单可靠等优点，因而在网络信息交流领域将会具有普遍应用价值，在网络的在线机译服务方面也会有广阔的市场。

说　明　书　附　图

符号序位		1	2	3	4	5	6	7	8	9	10	……
不同语言文字符号												
信息分类		备选信息项										
普通信息	概念定义：1、2、3 ……											
	时态信息：过去时、现在时…											
	语态信息：被动、使动…											
句法信息	句法成分：主语、谓词、宾语、修饰、补充、子句……											

图 1

序位	1	2	3	4	5	6	7	8	9	10	11
符号	我	*在*	银行	*附近*	看见	一个	带	望远镜	的	男孩	.
分类	信息项										
概念定义	自己	*表示时间处所* 2 范围 3 存在 4 参加	金融机构	*不远的地方*	看到	数字（省略"个"	？携带 ？附带 3 含有 4 引导 5 带动 6 长条物 7 区域	观察远距离物体的仪器	表修饰关系 2 表所有关系	男性未成年人	
时态					过去时						
语态											
句法成分	主词	谓词补充			谓词	宾词修饰				宾词	

图　2－1

I	saw	a	boy	with	a	telescope	near	the	bank	.
subject	predicate		object	? modification of the predicate ? modification of the object			? complementary of the predicate ? complementary of the object			

图 2 - 2

我	在	银行	附近	看见	一个	带	望远镜	的	男孩	.
						? 携带 ? 附带 3 含有 4 引导 5 带动				

图 2 - 3

"I see a boy with a telescope near the bank"的虚拟全息模板对话方法:

"**I see a boy** ***with a telescope*** 【" modification of the subject " **OR** " modification of the predicate " 】 **near the** ***bank*** " 【" institution for keeping or lending money " **OR** " edge of a river or canal " **OR** " long mound of earth /sand/snow " 】

图 2 - 4

修饰区			核心区			补充区		
				I				
主语区								
谓语区								
		? with a telescope		see		near the bank		
宾语区		? with a telescope		boy				
								标点
整句连词								

图 3

语种		英语	汉语	俄语	日语
基础词		laugh	笑	(基础词)	(基础词)
近义附码	A.程度 +	burst intolaughter guffaw	大笑 狂笑		
	B. 程度 -	smile 、 grin	微笑		
	C. 成语		笑逐言开		
	E. 书面语				
	F. 口语		笑哈哈、笑嘻嘻		
	G. 俚语				
	H. 俗语				
	I. 褒义				
	J. 贬义	snicker	傻笑		
	其它近义				

图 4

I	saw	a	boy	with	a	telescope	near	the	bank	.
INFORMATION COLUMN（将源语选项结果转换为目标语供译文用户进行直接查询）										
pronoun referring to the speaker	to sense with one's eyes	not a partiqular one	male child	using	not a particul-ar one	tube with a series of lenses for looking at very disdant objects	close/at only a little disdance in space or time	referring to a particular person or thing	institution for keeping or lending money	
singula			countable singula			singula			singula	
	past tense									
subject	predicate		object	modification of object			complementary of the predicate			

图 5